МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science»

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Слушатель Горшков Андрей Вячеславович

Москва, 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Композиционные материалы – это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т.е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита – железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Актуальность: Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

АТТЕСТАЦИОННОЕ ЗАДАНИЕ

1. Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER

https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing

2. Обучить алгоритм машинного обучения, который будет определять значения:

- Модуль упругости при растяжении, ГПа
- Прочность при растяжении, МПа

3. Написать нейронную сеть, которая будет рекомендовать:

- Соотношение матрица-наполнитель

4. Написать приложение, которое будет выдавать прогноз, полученный в задании 2 или 3 (один или два прогноза, на выбор учащегося)
5. Создать профиль на github.com
6. Сделать commit приложения на github.com
7. Сделать commit на веб-хостинг (По желанию учащегося)
8. Написать пояснительную записку к проекту, которая включает блок-схему и описание процесса подготовки, обучения моделей и инструкцию по установке и запуску приложения.

1 АНАЛИТИЧЕСКАЯ ЧАСТЬ

1.1 Предварительный анализ датасета

После объединения (по индексу) отдельных частей датасета итоговый датасет содержит 1023 строки и 13 столбцов – см. рисунок 1.

В результате визуального анализа выявлено, что первые 23 строки датасета содержат преимущественно целые и рациональные числа. При этом рациональные числа являются средними значениями первых 23 строк соответствующих столбцов датасета (за исключением показателя «Соотношение матрица-наполнитель»), что позволяет сделать предположение о том, что в этих строках имелись пропуски данных, которые были заполнены средними значениями первых 23 строк соответствующих столбцов.

D27 | 707,570887009741

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,% 2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
2	0	1,85714286	2030	738,736842	30	22,2678571	100	210	70	3000	220	0	4	57
3	1	1,85714286	2030	738,736842	50	23,75	284,615385	210	70	3000	220	0	4	60
4	2	1,85714286	2030	738,736842	49,9	33	284,615385	210	70	3000	220	0	4	70
5	3	1,85714286	2030	738,736842	129	21,25	300	210	70	3000	220	0	5	47
6	4	2,77133106	2030	753	111,86	22,2678571	284,615385	210	70	3000	220	0	5	57
7	5	2,76791809	2000	748	111,86	22,2678571	284,615385	210	70	3000	220	0	5	60
8	6	2,56962025	1910	807	111,86	22,2678571	284,615385	210	70	3000	220	0	5	70
9	7	2,56147541	1900	535	111,86	22,2678571	284,615385	380	75	1800	120	0	7	47
10	8	3,55701754	1930	889	129	21,25	300	380	75	1800	120	0	7	57
11	9	3,53233831	2100	1421	129	21,25	300	1010	78	2000	300	0	7	60
12	10	2,91967784	2160	933	129	21,25	300	1010	78	2000	300	0	7	70
13	11	2,87735849	1990	1628	129	21,25	300	1010	78	2000	300	0	9	47
14	12	1,59817352	1950	827	129	21,25	300	470	73,33333333	2455,555556	220	0	9	57
15	13	2,91967784	1980	568	129	21,25	300	470	73,33333333	2455,555556	220	0	9	60
16	14	4,02912621	1910	800	129	21,25	300	470	73,33333333	2455,555556	220	0	9	70
17	15	2,93478261	2030	302	129	21,25	300	210	70	3000	220	0	10	47
18	16	3,55701754	1880	313	129	21,25	300	210	70	3000	220	0	10	57
19	17	4,19354839	1950	506	129	21,25	300	380	75	1800	120	0	10	60
20	18	4,89795918	1890	540	129	21,25	300	380	75	1800	120	0	10	70
21	19	3,53233831	1980	1183	111,86	22,2678571	284,615385	1010	78	2000	300	0	0	0
22	20	2,87735849	2000	205	111,86	22,2678571	284,615385	1010	78	2000	300	90	4	47
23	21	1,59817352	1920	456	111,86	22,2678571	284,615385	470	73,33333333	2455,555556	220	90	4	57
24	22	4,02912621	1880	622	111,86	22,2678571	284,615385	470	73,33333333	2455,555556	220	90	4	60
25	23	2,58734764	1953,27493	1136,59613	137,6274196	22,3445336	234,716883	555,893453	80,80322176	2587,342983	246,613117	90	4	70
26	24	2,49991793	1942,59578	901,519947	146,2522078	23,0817575	351,231874	864,725484	76,17807508	3705,672523	226,22276	90	5	47
27	25	2,04647146	2037,63181	707,570887	101,6172513	23,1463928	312,307205	547,601219	73,81706662	2624,026407	178,198556	90	5	57
28	26	1,85647617	2018,22033	836,294382	135,4016966	26,4355146	327,510377	150,961449	77,21076158	2473,187195	123,344561	90	5	60
29	27	3,30553542	1917,90751	478,286247	105,7869296	17,8740999	328,154579	526,692159	72,34570879	3059,032991	275,57588	90	5	70
30	28	2,70955409	1892,07112	641,052549	96,56329319	22,9892906	262,956722	804,592621	74,51135922	2288,967377	126,816339	90	7	47
31	29	2,28282531	2008,35759	393,967325	149,3728324	21,6617507	330,498641	535,371459	72,24492408	2704,445081	261,077072	90	7	57
32	30	1,97814017	1973,6291	991,724095	149,3721279	19,7505779	232,058191	485,453778	75,66570056	2448,943079	162,493694	90	7	60
33	31	1,77143639	1872,49156	801,033883	79,79454787	22,2963037	340,736898	864,929184	70,94759156	2796,785402	123,356264	90	7	70
34	32	3,27708699	2010,04701	339,550423	67,49899306	24,280609	254,949084	117,535234	67,47870663	2462,605386	207,018581	90	9	47
35	33	2,98436223	1912,31544	1183,09185	133,5490007	23,2637966	314,996125	377,389009	75,29045222	2303,770656	200,580249	90	9	57
36	34	2,91614962	1879,96985	1003,27018	109,2395305	25,6827595	294,048537	408,354239	71,70085562	3086,546196	192,191162	90	9	60

Рисунок 1 – Фрагмент датасета

Остальные 1000 строк датасета содержат 14-значные и 15-значные иррациональные числа без пропусков данных. Установлено, что средние значения и среднеквадратичные отклонения (СКО) значений столбцов этой

части датасета с высокой точностью совпадают со средними значениями и СКО соответствующих столбцов, как первых 23 строк датасета, так и полного датасета. С высокой степенью уверенности можно утверждать, что 14-значные и 15-значные иррациональные числа в датасете являются результатом вычислений, а не результатом прямых или косвенных измерений (маловероятно, что средства измерений физических характеристик композитов позволяют проводить измерения с такой высокой точностью).

На основании выявленной информации можно предположить, что первые 23 строки датасета являются планом многофакторного измерительного или численного эксперимента, пропуски данных в котором были заполнены средними значениями первых 23 строк соответствующих столбцов. Следовательно, данные в этих строках являются достоверными (с некоторой точностью) значениями. Для определенности та часть датасета, которая состоит из первых 23 строк, в данной работе называется базовым датасетом.

Также можно предположить, что последние 1000 строк датасета представляют собой совокупность двух видов значений данных.

Первый вид значений – значения в промежуточных точках базового датасета, вычисленные по аппроксимациям зависимостей переменных базового датасета.

Второй вид значений – гауссовский шум, сгенерированный по средним значениям и СКО данных в соответствующих столбцах базового датасета.

Возможный третий вид значений – суперпозиция первого и второго вида значений в данной работе не рассматривается.

В итоге проведенного предварительного анализа принято следующее решение – построить математические модели целевых переменных на данных каждого из трех следующих вариантов набора данных:

1) базовый датасет (первые 23 строки);

2) полный датасет (1023 строки);

3) очищенный датасет, полученный путем фильтрации полного датасета от помех и шума (N строк: $N < 1023$).

1.2 Последовательность операций (пайплайн) машинного обучения

1.2.1 Пайплайн машинного обучения на данных базового датасета

Принятый в данной работе пайплайн машинного обучения на данных базового датасета состоит из следующих операций – см. рисунок 2:

1) разведочный анализ данных;

2) отбор значимых признаков;

3) обучение различных моделей на всех данных (23 строки) базового датасета;

4) вычисление коэффициента детерминации R^2 и других метрик моделей, вычисление доверительного интервала для расчетных значений лучшей модели.

Особенностью данного пайплайна является отсутствие таких процедур, как удаление выбросов и разделение базового датасета на обучающую и тестовую выборки. Основанием для этого является малое число объектов (точек) базового датасета.

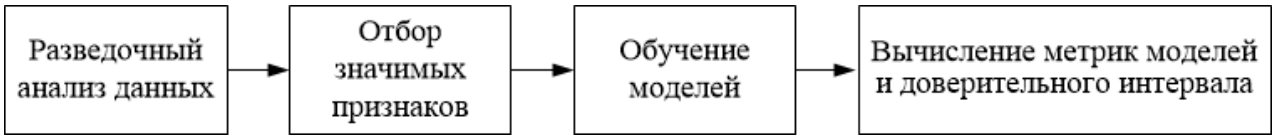


Рисунок 2 – Пайплайн обработки базового датасета

1.2.2 Пайплайн машинного обучения на данных полного датасета

Принятый в данной работе пайплайн машинного обучения на данных полного датасета состоит из следующих операций – см. рисунок 3:

1) разведочный анализ данных;

2) удаление выбросов;

3) нормализация данных (приведение к нормальному закону распределения);

4) отбор значимых признаков;

5) разбиение на обучающую (train) и тестовую (test) выборки;

6) масштабирование признаков;

7) обучение различных моделей на всех данных полного датасета;

8) вычисление коэффициента детерминации R^2 и других метрик моделей, вычисление доверительного интервала для прогноза значений лучшей модели.

Для исключения путаницы в терминологии в данной работе под нормализацией данных понимается процедура приведения распределения данных к закону распределения, близкому к нормальному.

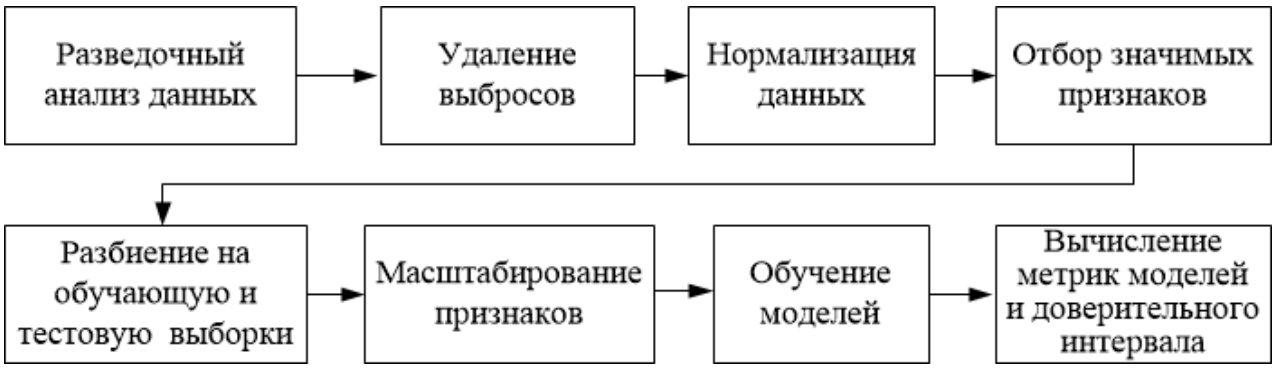


Рисунок 3 – Пайплайн обработки полного датасета

1.2.3 Пайплайн машинного обучения на данных очищенного датасета

Принятый в данной работе пайплайн машинного обучения на данных очищенного датасета состоит из следующих операций – см. рисунок 4:

1) разведочный анализ данных

2) очистка (фильтрация) полного датасета от помех и шума;

3) нормализация данных (приведение к нормальному закону распределения);

4) отбор значимых признаков;

5) разбиение на обучающую (train) и тестовую (test) выборки;

6) масштабирование признаков;

7) обучение различных моделей на всех данных полного датасета;

8) вычисление коэффициента детерминации R^2 и других метрик моделей, вычисление доверительного интервала для прогноза значений лучшей модели.

Данный пайплайн отличается от пайплайна обучения на данных полного датасета только заменой процедуры удаления выбросов процедурой фильтрации датасета от шума и помех.

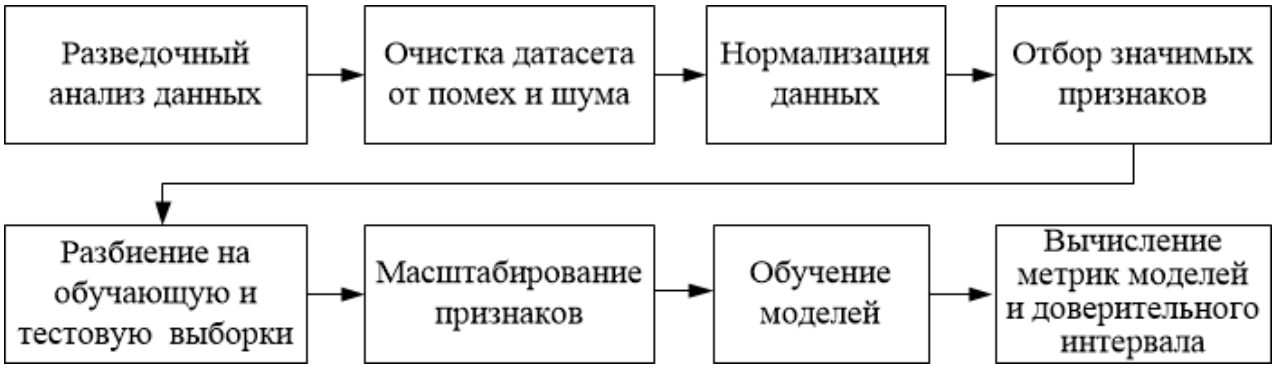


Рисунок 4 – Пайплайн обработки очищенного датасета

1.3 Методы, используемые для выполнения операций пайплайнов

1.3.1 Разведочный анализ данных

При разведочном анализе данных выполнялись следующие процедуры – визуализация данных и вычисление статистических характеристик датасета.

В рамках визуализации данных проводилось построение гистограмм распределения каждой из переменной, попарных графиков рассеяния точек, диаграмм boxplot («ящиков с усами»).

Кроме того, вычислены основные статистические характеристики всех целевых переменных и признаков датасета:

- средние значения;
- медианные значения;
- среднеквадратичные отклонения;
- другие статистические параметры.

1.3.2 Удаление выбросов в данных

Для поиска выбросов использовались следующие методы:

- метод трех среднеквадратичных отклонений (СКО) σ – обнаружение данных, значения которых отклоняются от средних значений более, чем на 3σ;

- метод межквартильного размаха – обнаружение данных, значения которых расположены за пределами третьего (Q3) и первого (Q1) квартилями;

- метод Isolation Forest – алгоритм для обнаружения аномалий в данных (выбросов), использующий решающие бинарные деревья [1]. Идея алгоритма заключается в том, что при разделении пространства данных линиями, ортогональными началу координат, те точки данных, для изоляции которых от датасета требуется меньше разделений, с высокой вероятностью являются аномалиями данных (особенно точки, являющиеся листьями).

1.3.3 Нормализация данных

В полном датасете имеются нулевые значения некоторых признаков. По этой причине для нормализации данных использовалось преобразование Йео – Джонсона [2], позволяющее трансформировать нулевые и отрицательные значения.

1.3.4 Отбор значимых признаков

Для отбора значимых признаков использовались следующие операции:

- вычисление коэффициентов корреляции Пирсона и Спирмена;
- вычисление статистических показателей значимости признаков – взаимной информации MI (mutual information) и F-теста;
- анализ мультиколлинеарности признаков путем вычисления коэффициента увеличения дисперсии VIF (variance inflation factor);
- пошаговый отбор признаков (stepwise regression).

1.3.5 Разбиение на обучающую и тестовую выборки

Разбиение данных на обучающую (train) и тестовую (test) выборки проводилось в соотношении 70/30.

Для предотвращения пере- и недообучения моделей (с целью обеспечения их приемлемой обобщающей способности) проводился контроль подобия обучающей и тестовой выборок. Контроль проводился путем сравнения значений аналитических коэффициентов детерминации R^2 для обучающей и тестовой выборок.

Аналитический коэффициент детерминации является коэффициентом детерминации для линейной модели, построенной методом наименьших квадратов (МНК), и вычисляется по значениям коэффициентов корреляции выборок следующим образом [3]

$$R^2 = 1 - \frac{detA^+}{detA}, \quad (1)$$

где $detA^+$ – определитель матрицы парных коэффициентов корреляции между целевой переменной и признаками;

$detA$ – определитель матрицы парных коэффициентов межфакторной корреляции (между признаками).

Предполагается, что если обучающая и тестовая выборки подобны (то есть имеют сравнимые коэффициенты корреляции и, следовательно, сравнимые коэффициенты детерминации), то это будет способствовать предотвращению пере- и недообучения моделей.

В данной работе при существенном – на порядок и более – различии аналитических коэффициентов детерминации обучающей и тестовой выборок данное разбиение отбрасывалось и для обучения моделей использовалось другое разбиение, при котором аналитические коэффициенты детерминации R^2 для обучающей и тестовой выборок были сравнимы.

Необходимо отметить еще одно возможное назначение аналитического коэффициента детерминации R^2 для МНК – его использование для оценки принципиальной возможности построения неконстантной (нетривиальной) модели для данного датасета. Предполагается, что если вычисленное по выражению (1) значение аналитического коэффициента детерминации R^2 для обучающей выборки очень мало ($R^2 \to 0$), то построить неконстантную, то есть пригодную для практического применения, математическую модель с обобщающей способностью на этих данных практически невозможно или маловероятно для любого алгоритма.

1.3.6 Масштабирование признаков

Масштабирование признаков применялось для улучшения сходимости алгоритмов обучения различных моделей. В качестве методов масштабирования использовались:

- стандартизация (центрирование данных относительно их среднего значения и приведение к их среднеквадратичному отклонению);

- приведение к межквартильному диапазону

$$x_i = \frac{x_i - Q1}{Q3 - Q1}, \tag{2}$$

где $Q1$ и $Q3$ – первый и третий квартиль распределения данных, соответственно.

1.3.7 Машинное обучение моделей

Машинное обучение математических моделей в данной работе выполнено на языке программирования Python 3.

1.3.7.1 Целевые переменные «Модуль упругости при растяжении» и «Прочность при растяжении»

Для построения математических моделей целевых переменных «Модуль упругости при растяжении» и «Прочность при растяжении» применялись следующие алгоритмы из библиотеки scikit-learn языка программирования Python 3.

а) Параметрические алгоритмы регрессии:

- LinearRegression;
- ElasticNet.

б) Ансамблевые алгоритмы:

- GradientBoostingRegressor;
- RandomForestRegressor.

в) Непараметрические алгоритмы регрессии:

- KneighborsRegressor;
- TheilSenRegressor [4].

Для построения моделей с наилучшими характеристиками проводился поиск оптимальных значений их гиперпараметров по сетке с перекрестной проверкой с количеством блоков, равным 10.

1.3.7.2 Целевая переменная «Соотношение матрица – наполнитель»

Для построения математических моделей целевой переменной «Соотношение матрица – наполнитель» рассматривались полносвязные нейронные сети с двумя скрытыми слоями и одним выходным слоем.

1.3.8 Вычисление метрик обученных моделей и доверительных интервалов для прогноза значений целевых переменных

Для анализа точности и обобщающей способности обученных моделей вычислялись метрики R^2, MAE, RMSE для обучающих и тестовых выборок. Для наглядности метрики обученных моделей отражены в сводных таблицах метрик разных моделей.

Согласно вычисленным статистическим характеристикам полного датасета все целевые переменные и признаки имеют сравнительно большую дисперсию. Следовательно, доверительные интервалы целевых переменных также будут иметь относительно большие значения, вследствие прямой пропорциональности ширины доверительного интервала от дисперсии значений переменных. Учитывая это обстоятельство принято решение рассматривать в данной работе не только точечные прогнозируемые значения целевых переменных, но и их доверительные интервалы с уровнем значимости 0,05. Разумеется, что представление результата прогноза в виде доверительного интервала позволит достоверно, то есть с определенной точностью, прогнозировать целевые переменные по обученным моделям.

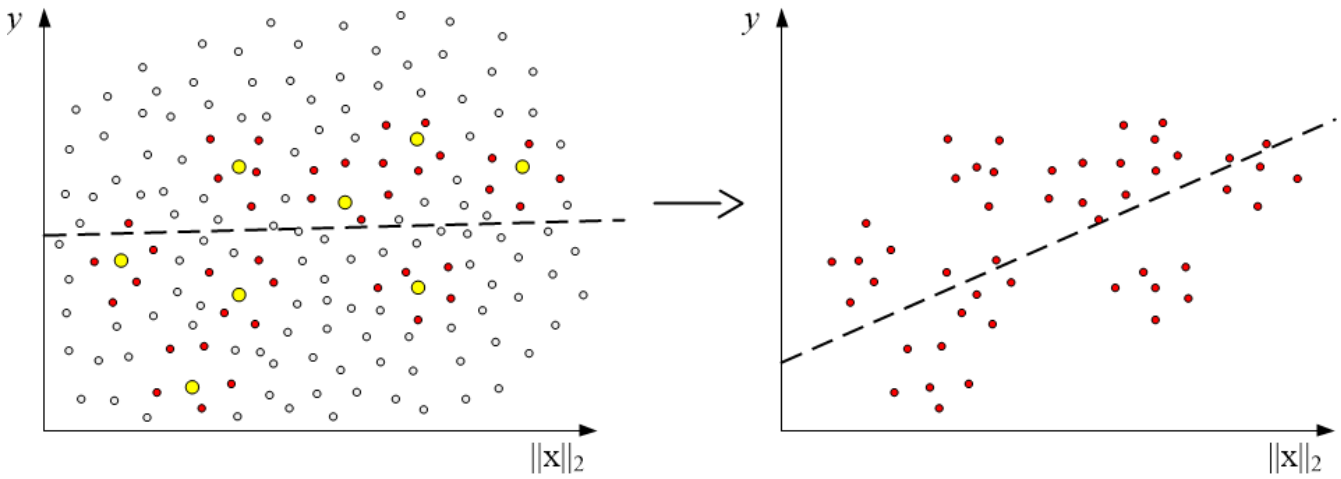
1.4 Фильтрация полного датасета от помех и шума

Данная процедура имеет первостепенное значение для решения поставленной задачи – создания неконстантных, то есть пригодных для

практического применения, математических моделей целевых переменных на основе предоставленного датасета.

В основе идеи фильтра лежит предположение, что первые 23 строки полного датасета являются планом многофакторного эксперимента. Следовательно, данные в этих строках являются достоверными, разумеется, с некоторой точностью, которая определяется некоторым доверительным интервалом. Тогда те значения данных, которые находятся в этом доверительном интервале, с высокой вероятностью также являются достоверными значениями. И наоборот, те данные, которые расположены вне доверительного интервала, с высокой вероятностью содержат значительные составляющие помех и (или) шумов и потому должны быть исключены из процесса обучения моделей.

Таким образом, объекты базового датасета можно рассматривать как центроиды, которые позволяют отобрать достоверные значения, например, методом «ближайших соседей» – см. рисунок 5.



а) полный датасет; б) очищенный датасет (k_{nn} = 5)

Рисунок 5 – Очистка датасета от помех и шума с помощью фильтра, ядром которого являются объекты (выделены желтым цветом) базового датасета.
Примечание. Пунктиром показаны линии регрессии

При этом предполагаем, что центроиды и «ближайшие соседи» имеют одинаковую значимость (вес), так как нет оснований полагать, что данные

базового датасета являются более достоверными, чем отобранные в результате фильтрации объекты. Следовательно, для дальнейшего машинного обучения моделей центроиды и «ближайшие соседи» являются равнозначными данными.

Для улучшения процедуры фильтрации можно использовать итеративный алгоритм фильтра – сначала отбираются «ближайшие соседи» для объектов базового датасета, обучается модель регрессии, затем отбираются «ближайшие соседи» уже в некотором доверительном интервале для всей кривой регрессии, снова обучается модель и т. д., пока полученная модель регрессии не станет устойчивой. В данной работе рассматривается только первая итерация алгоритма фильтра – отбираются «ближайшие соседи» только для объектов базового датасета.

Необходимо отметить, что если все последние 1000 строк полного датасета сгенерированы случайным образом, то эту процедуру следует рассматривать не как фильтрацию датасета, а как размножение (oversampling) достоверных данных – объектов базового датасета.

2 ПРЕДОБРАБОТКА ДАННЫХ

Программный код и результаты предобработки данных приведены в следующих файлах, расположенных на странице автора в репозитории GitHub – см. 4.2:

first_23.ipynb

y_1_full.ipynb

y_1_filtered.ipynb

y_2_full.ipynb

y_2_filtered.ipynb

y_3_full_NN.ipynb

y_3_filtered_NN.ipynb

y_3_filtered_Regressor.ipynb

В программном коде и пояснительной записке приняты следующие обозначения целевых переменных.

y_1 – 'Модуль упругости при растяжении, ГПа'

y_2 – 'Прочность при растяжении, МПа'

y_3 – 'Соотношение матрица-наполнитель'

2.1 Визуализация данных

Визуализация данных проводилась с использованием библиотек matplotlib и seaborn языка программирования Python 3.

На рисунках 6, 7, 9, 10 приведены точечные диаграммы и гистограммы распределений целевых переменных и признаков для базового, полного и очищенных датасетов.

На рисунке 8 приведен график boxplot для целевых переменных и признаков полного датасета после удаления выбросов методом Isolation Forest.

Визуальный анализ приведенных на рисунке 7 распределений позволяет сделать вывод, что полный датасет содержит большое количество данных, являющихся шумом или помехами, в результате чего корреляция между целевыми переменными и признаками практически отсутствует.

2.1.1 Визуализация данных базового датасета

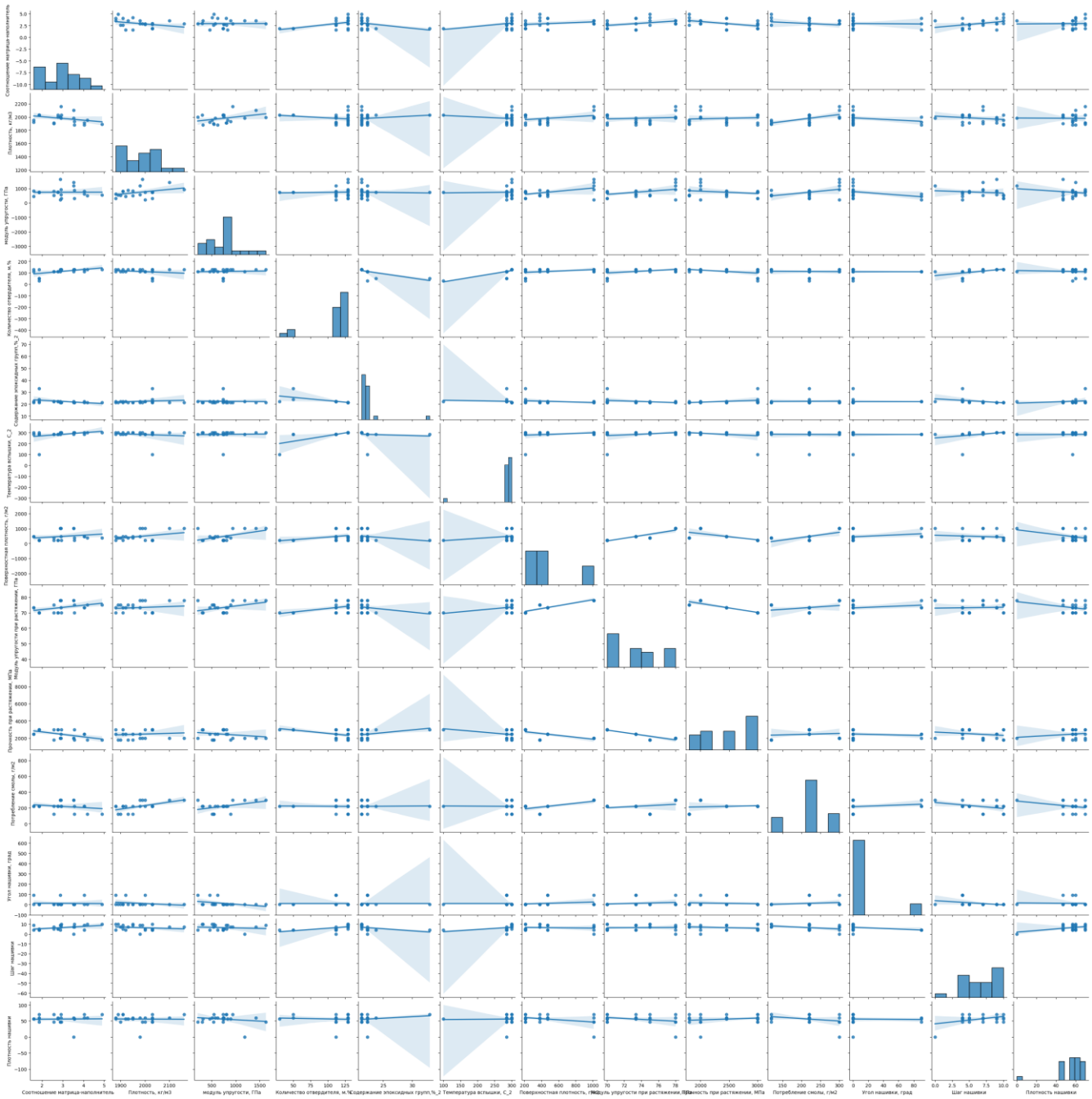


Рисунок 6 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков базового датасета

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

2.1.2 Визуализация данных полного датасета

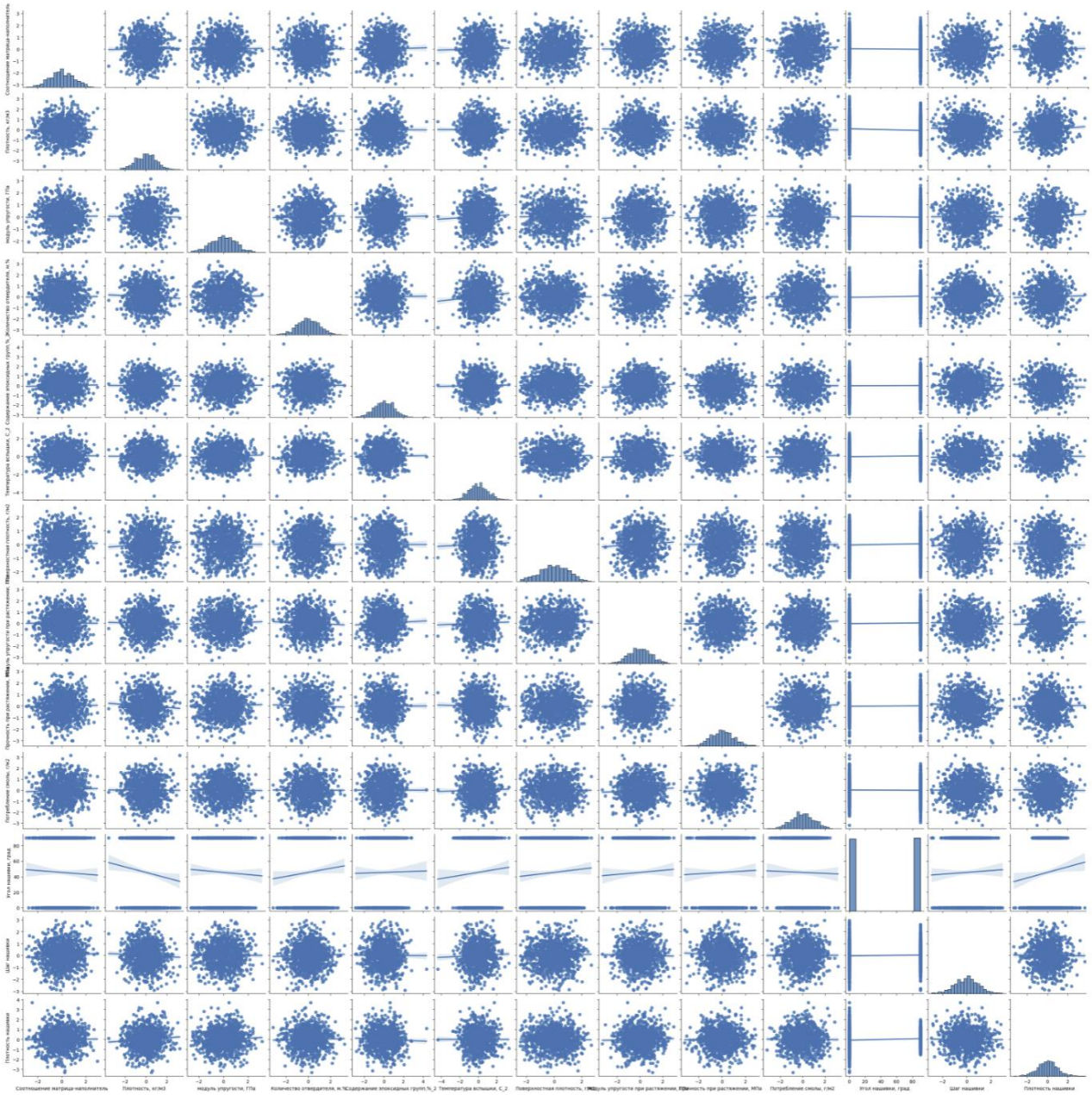


Рисунок 7 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков полного датасета (после удаления выбросов и преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

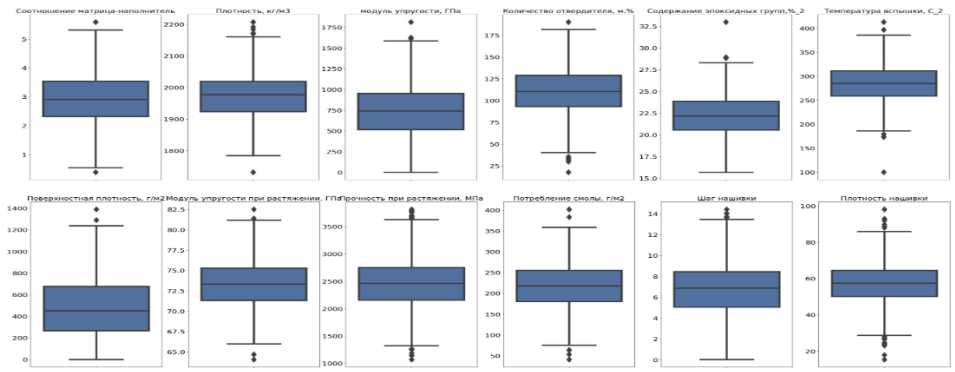


Рисунок 8 – Графики boxplot для целевых переменных и признаков полного датасета (после удаления выбросов методом Isolation Forest)

2.1.3 Визуализация данных очищенных датасетов

а) Число «ближайших соседей» k_{nn} = 3

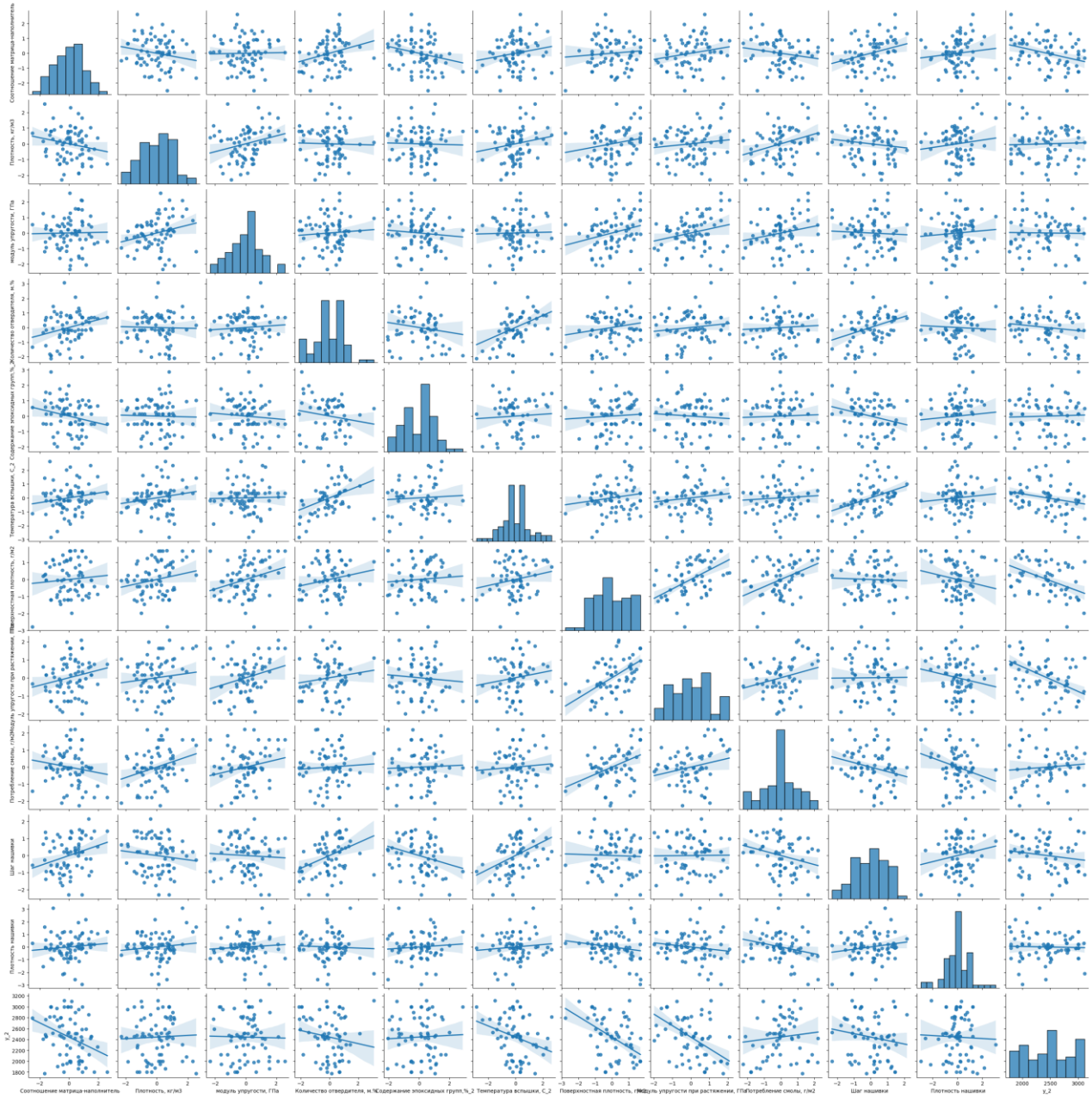


Рисунок 9 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков очищенного датасета (k_{nn} = 3; после преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

б) Число «ближайших соседей» k_{nn} = 5

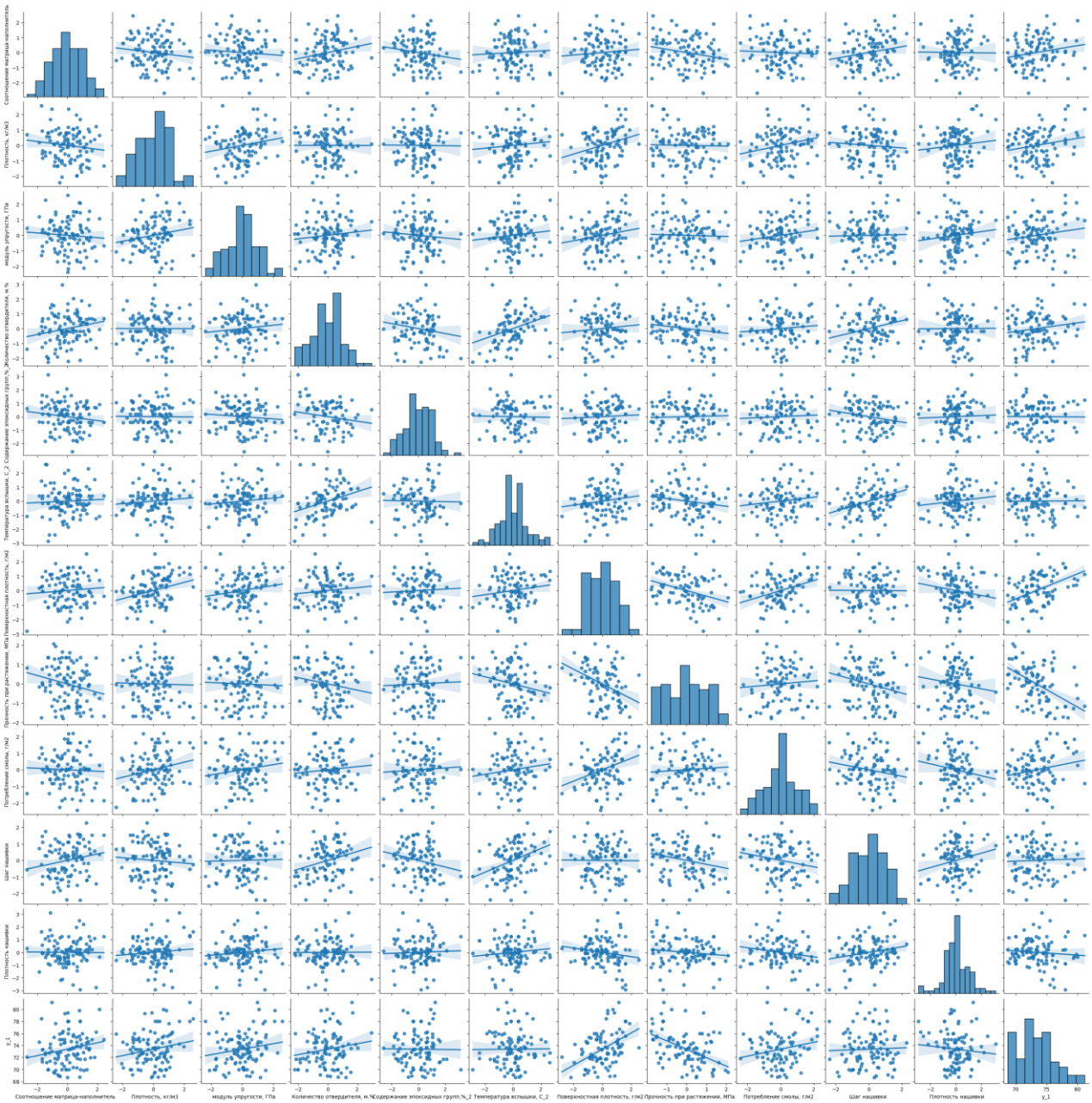


Рисунок 10 – Точечные диаграммы и гистограммы распределений целевых переменных и признаков очищенного датасета (k_{nn} = 5; после преобразования Йео-Джонсона)

Примечание. На диаграммах показаны линии регрессии и их доверительные интервалы

2.2 Коэффициенты корреляции и аналитический коэффициент детерминации для МНК

На рисунках 11 – 18 приведены матрицы попарных коэффициентов корреляции Пирсона для целевых переменных и признаков базового, полного и очищенного датасета.

В таблицах 1 – 3 приведены расчетные значения аналитических коэффициентов детерминации R^2 для целевых переменных базового, полного и очищенных датасетов. Аналитические коэффициенты детерминации рассчитаны как для целых датасетов (до их разбиения на обучающую и тестовые выборки), так и для обучающих и тестовых выборок.

Согласно приведенным в таблице 1 данным для базового датасета значения аналитического коэффициента детерминации $R^2 = 1$ для целевых переменных 'Модуль упругости при растяжении, ГПа' и 'Прочность при растяжении, МПа'. Следовательно, данные переменные имеют линейные функциональные зависимости с некоторыми признаками. Целевая переменная 'Соотношение матрица-наполнитель' имеет значение аналитического коэффициента детерминации $R^2 = 0{,}393$, что позволяет сделать вывод о принципиальной возможности построения неконстантной модели для данной целевой переменной по объектам базового датасета.

Согласно приведенным в таблице 2 данным для полного датасета все целевые переменные имеют практически нулевые значения аналитического коэффициента детерминации R^2, что позволяет сделать вывод о принципиальной невозможности или малой вероятности построения неконстантных математических моделей для целевых переменных по объектам полного датасета.

Согласно приведенным в таблице 3 данным для очищенных датасетов (при $k_{nn} = 3$ и $k_{nn} = 5$) целевые переменные имеют значения аналитического коэффициента детерминации R^2 в диапазоне от 0,242 до 0,812, что позволяет сделать вывод о принципиальной возможности построения неконстантных моделей для целевых переменных по объектам очищенных датасетов.

2.2.1 Коэффициенты корреляции и детерминации для базового датасета

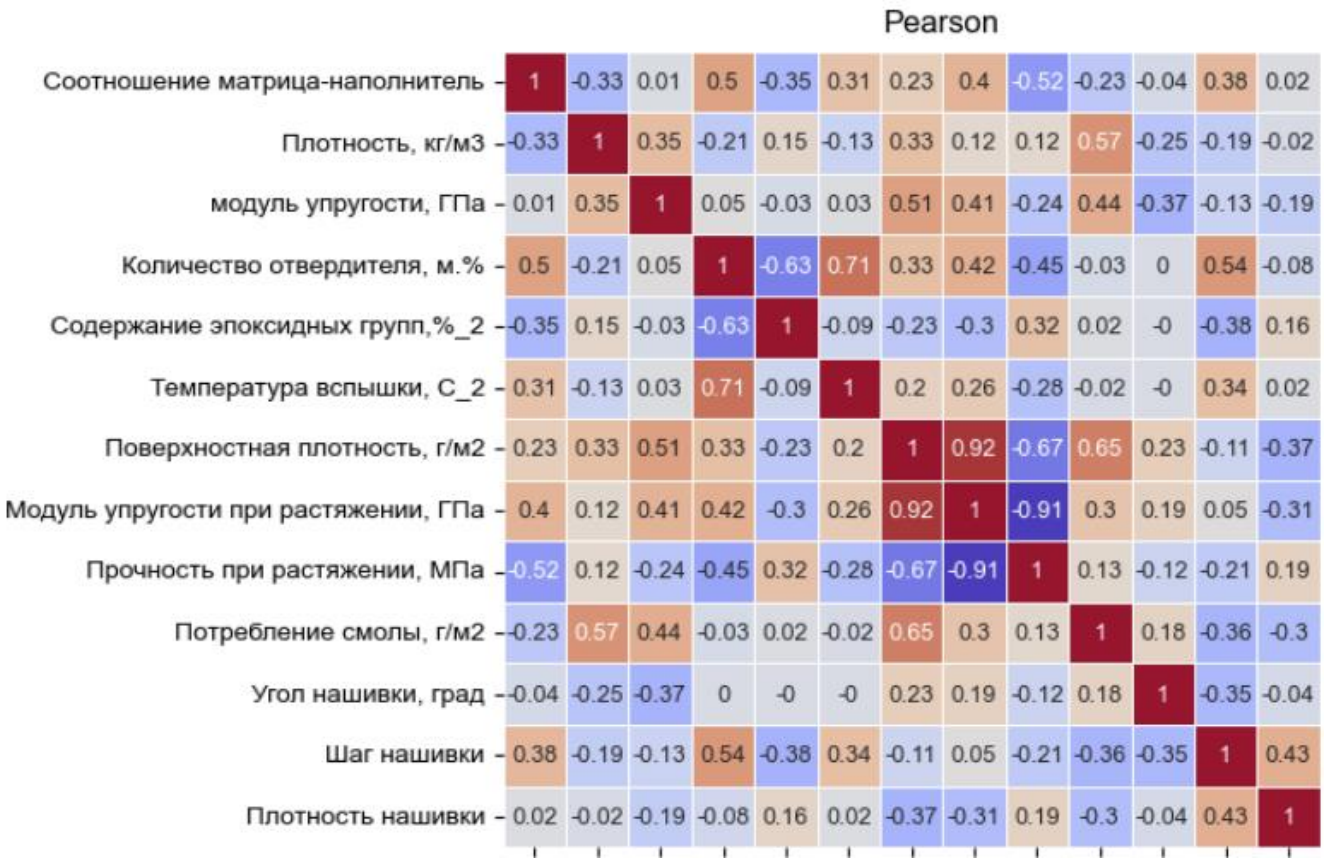


Рисунок 11 – Матрица корреляции признаков базового датасета

Рисунок 12 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (базовый датасет)

Таблица 1 – Аналитические коэффициенты детерминации для целевых переменных (базовый датасет)

	Аналитический коэффициент детерминации для МНК R^2		
	до разбиения на train- и test-выборки	train-выборка	test-выборка
'Модуль упругости при растяжении'	1	разбиение не проводилось	разбиение не проводилось
'Прочность при растяжении'	1	разбиение не проводилось	разбиение не проводилось
'Соотношение матрица - наполнитель'	0,393	разбиение не проводилось	разбиение не проводилось

2.2.2 Коэффициенты корреляции и детерминации для полного датасета

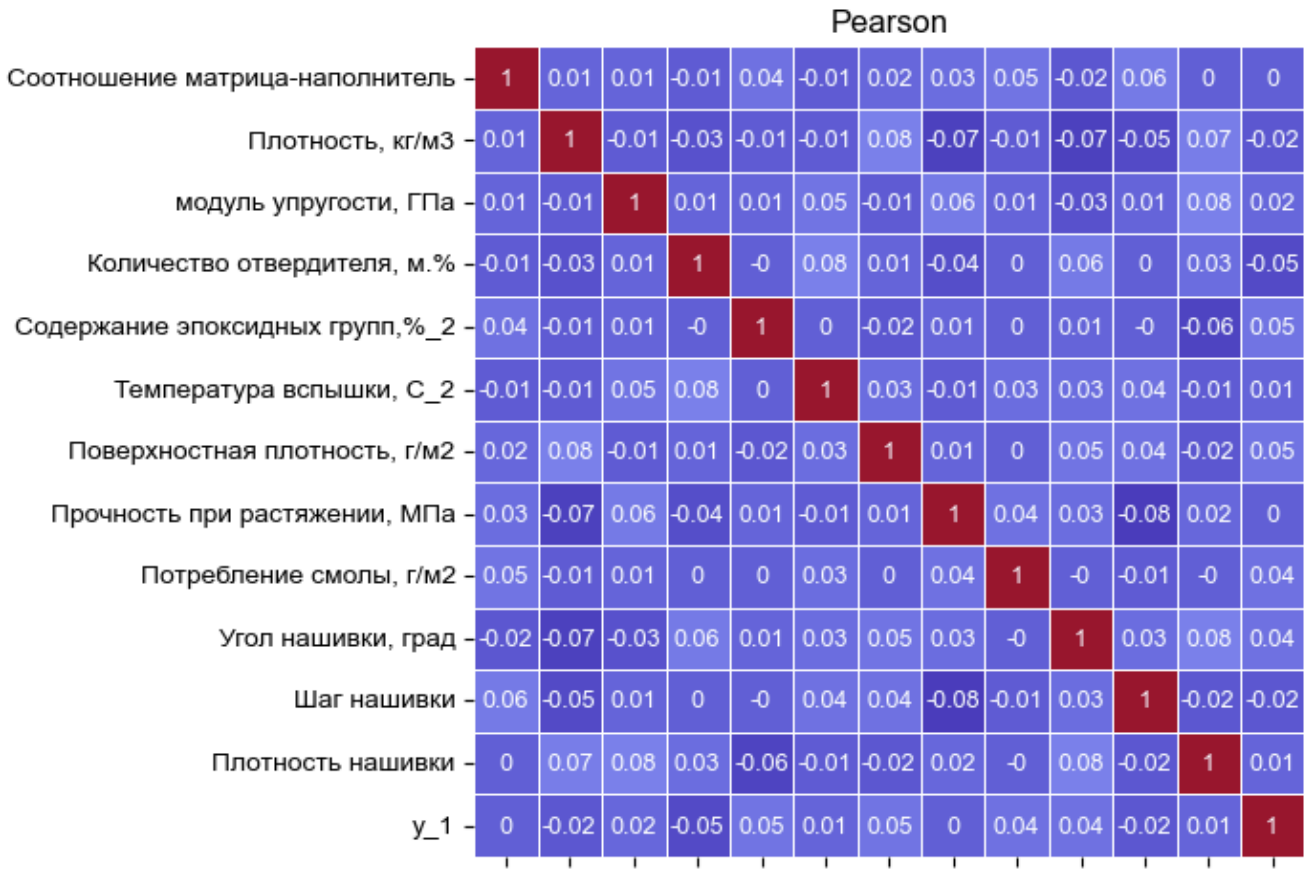


Рисунок 13 – Матрица корреляции признаков полного датасета

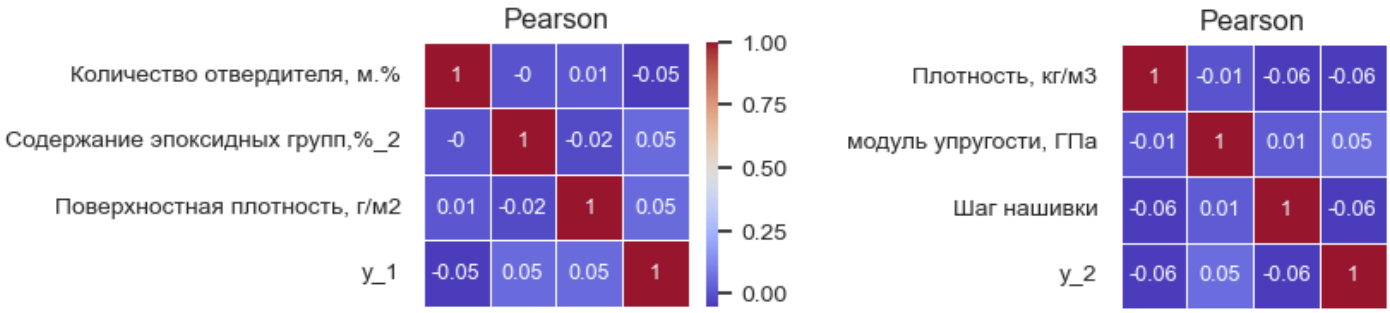


Рисунок 14 – Матрицы корреляции отобранных признаков для переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' (полный датасет)

Рисунок 15 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (полный датасет)

Таблица 2 – Аналитические коэффициенты детерминации для целевых переменных (полный датасет)

	Аналитический коэффициент детерминации для МНК R^2		
	до разбиения на train- и test-выборки	train-выборка	test-выборка
'Модуль упругости при растяжении'	0,012	0,008	0,005
'Прочность при растяжении'	0,016	0,009	0,017
'Соотношение матрица - наполнитель'	0,012	0,008	0,057

2.2.3 Коэффициенты корреляции и детерминации для очищенного датасета

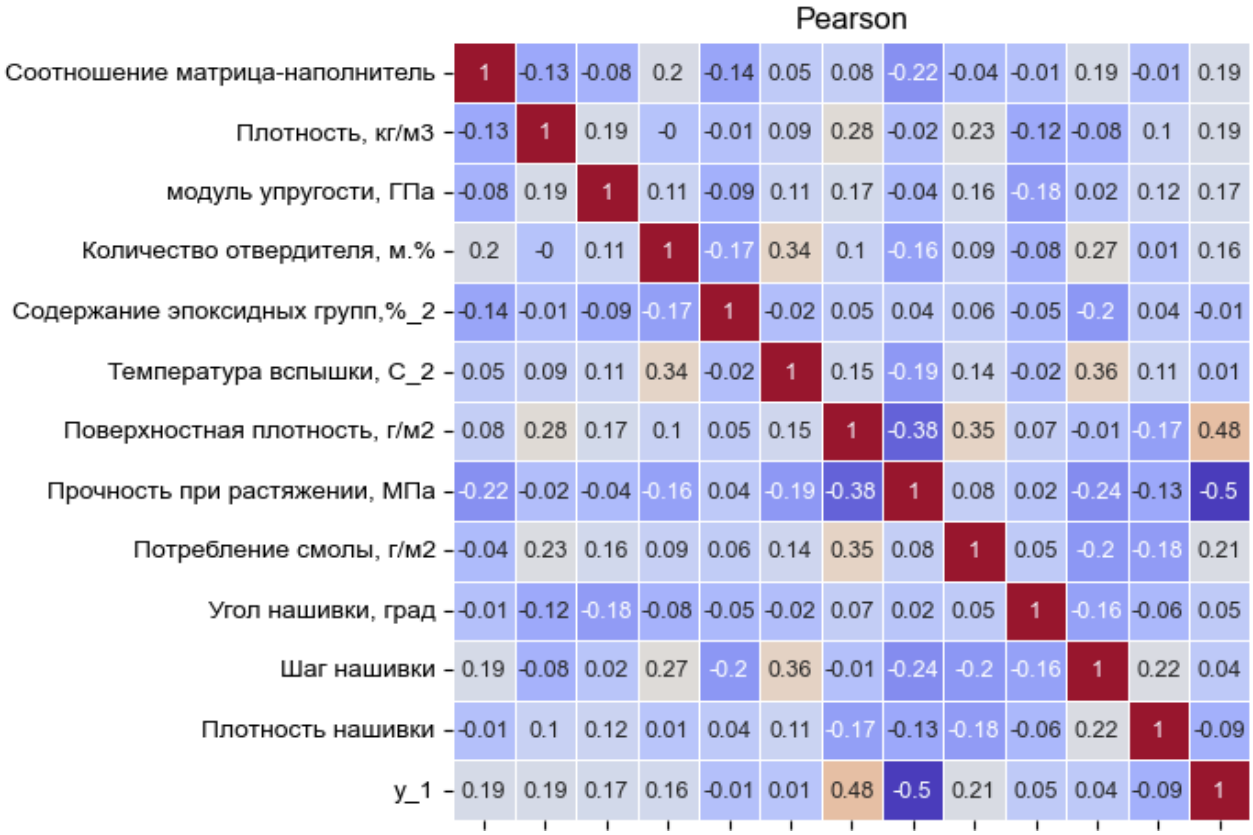


Рисунок 16 – Матрица корреляции признаков очищенного датасета (k_{nn} = 5)

Рисунок 17 – Матрицы корреляции отобранных признаков для переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' (очищенные датасеты; k_{nn} = 5 и k_{nn} = 3, соответственно)

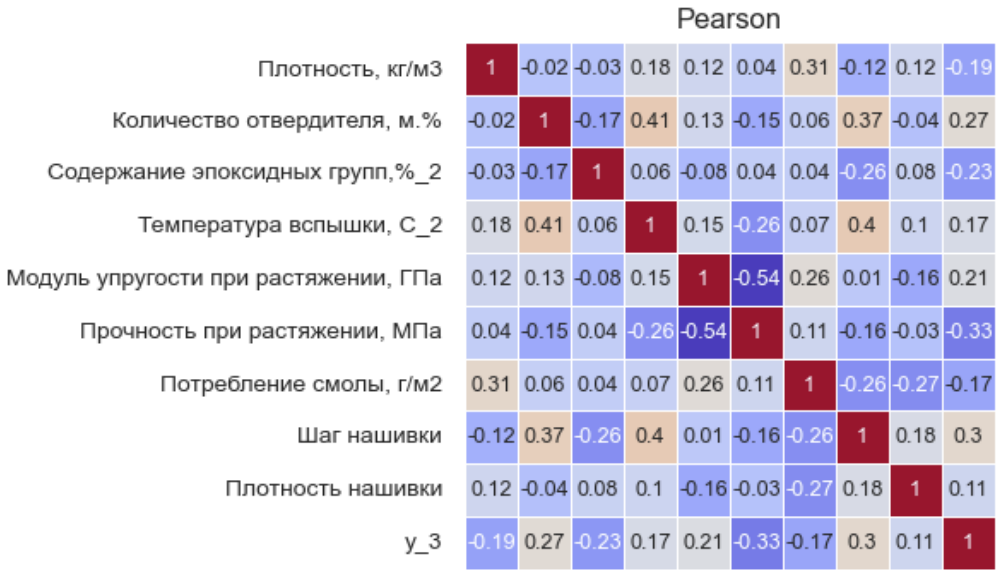


Рисунок 18 – Матрица корреляции отобранных признаков для переменной 'Соотношение матрица - наполнитель' (очищенный датасет; k_{nn} = 3)

Таблица 3 – Аналитические коэффициенты детерминации для целевых переменных (очищенный датасет)

	Аналитический коэффициент детерминации для МНК R^2		
	до разбиения на train- и test-выборки	train-выборка	test-выборка
'Модуль упругости при растяжении'	0,434	0,361	0,512
'Прочность при растяжении'	0,523	0,397	0,812
'Соотношение матрица - наполнитель'	0,265	0,242	0,482

3 РЕЗУЛЬТАТЫ МАШИННОГО ОБУЧЕНИЯ МОДЕЛЕЙ

Программный код и результаты машинного обучения моделей приведены в следующих файлах, расположенных на странице автора в репозитории GitHub – см. 4.2:

first_23.ipynb
y_1_full.ipynb
y_1_filtered.ipynb
y_2_full.ipynb
y_2_filtered.ipynb
y_3_full_NN.ipynb
y_3_filtered_NN.ipynb
y_3_filtered_Regressor.ipynb

3.1 Результаты обучения моделей на данных базового датасета

В результате машинного обучения модели LinearRegression получены выражения (3, 4) для линейных функциональных зависимостей целевых переменных 'Модуль упругости при растяжении, ГПа' и 'Прочность при растяжении, МПа' от признаков 'Поверхностная плотность, г/м2' и 'Потребление смолы, г/м2'. Полученные выражения подтверждают полученные ранее результаты расчета аналитического коэффициента детерминации $R^2 = 1$ для МНК – см. 2.2 и таблицу 1.

В таблице 4 приведены метрики для обученных моделей различных регрессоров для переменной 'Соотношение матрица - наполнитель'. Максимальное значение коэффициента детерминации $R^2 = 0{,}505$ для обучающей выборки (все значения базового датасета) имеет модель GradientBoostingRegressor.

Следует отметить, что значение коэффициента детерминации $R^2 = 0{,}393$ для модели LinearRegression идентично значению аналитического коэффициента детерминации R^2 для МНК – см. 2.2 и таблицу 1.

Доверительный интервал для расчета значений обучающей выборки переменной 'Соотношение матрица - наполнитель' по модели GradientBoostingRegressor приведен на рисунке 19. Все значения обучающей выборки расположены в пределах доверительного интервала, что позволяет сделать вывод о приемлемой точности обученной модели.

3.1.1 'Модуль упругости при растяжении'

$$y\_1 = 73{,}51 + 0{,}0128 \cdot x\_6 - 0{,}0282 \cdot x\_9 , \quad (3)$$

где x_6 = 'Поверхностная плотность, г/м2';

x_9 = 'Потребление смолы, г/м2'.

3.1.2 'Прочность при растяжении'

$$y\_2 = 1582{,}9 - 2{,}094 \cdot x\_6 + 8{,}44 \cdot x\_9 , \quad (4)$$

где x_6 = 'Поверхностная плотность, г/м2';

x_9 = 'Потребление смолы, г/м2'.

3.1.3 'Соотношение матрица - наполнитель'

Таблица 4 – Метрики различных моделей для переменной 'Соотношение матрица - наполнитель' (базовый датасет)

	LinearRegression	ElasticNet	GradientBoosting	RandomForest	KNeighbors	TheilSen
RMSE	0.687	0.688	0.620	0.675	0.657	0.900
MAE	0.518	0.521	0.438	0.512	0.469	0.700
R2	0.393	0.391	0.505	0.414	0.445	-0.042
Adj.R2	0.332	0.331	0.456	0.355	0.389	-0.146

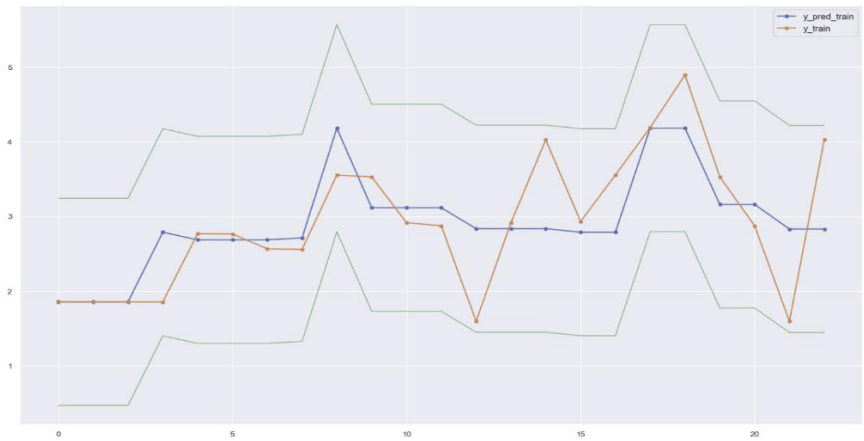


Рисунок 19 – Доверительный интервал для расчета обучающих значений переменной 'Соотношение матрица - наполнитель' (базовый датасет; модель – GradientBoostingRegressor)

Примечание. Зеленым цветом показаны границы доверительного интервала

3.2 Результаты обучения моделей на данных полного датасета

В таблице 5 приведены метрики для различных обученных моделей для переменной 'Модуль упругости при растяжении'. Максимальное значение коэффициента детерминации R^2 = 0,01 для тестовой выборки имеет модель KneighborsRegressor.

В таблице 6 приведены метрики для различных обученных моделей для переменной 'Прочность при растяжении'. Максимальное значение коэффициента детерминации R^2 = 0,012 для тестовой выборки имеет модель TheilSenRegressor.

Следует отметить, что полученные значения коэффициентов детерминации R^2 для обучающих значений модели LinearRegression идентичны соответствующим значениям аналитического коэффициента детерминации R^2 для МНК – см. таблицу 2.

На рисунке 20 приведены метрики для обученной модели искусственной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для переменной 'Соотношение матрица - наполнитель'. Значение коэффициента детерминации для тестовой выборки составило R^2 = 0,026.

Доверительный интервал для прогноза тестовых значений обучающей выборки переменной 'Соотношение матрица - наполнитель' приведен на рисунке 21. Практически все значения тестовой выборки расположены в пределах доверительного интервала.

Полученные метрики моделей, обученных на данных полного датасета, позволяют сделать вывод о том, что на этих данных **не удалось** построить неконстантные, то есть пригодные для практического применения, модели для целевых переменных. Данный вывод подтверждает, сделанный в 2.2 вывод о принципиальной невозможности или малой вероятности построения неконстантных моделей для целевых переменных по объектам полного датасета.

3.2.1 'Модуль упругости при растяжении'

Таблица 5 – Метрики различных моделей для переменной 'Модуль упругости при растяжении' (полный датасет)

	LinearRegression	ElasticNet	GradientBoosting	RandomForest	KNeighbors	TheilSen
RMSE train	3.015	3.017	2.927	2.996	3.012	3.016
RMSE test	2.957	2.955	2.958	2.955	2.946	2.958
MAE train	2.438	2.441	2.359	2.424	2.442	2.436
MAE test	2.361	2.357	2.354	2.356	2.342	2.358
R2 train	0.008	0.006	0.065	0.020	0.010	0.007
R2 test	0.003	0.004	0.002	0.005	0.010	0.002

3.2.2 'Прочность при растяжении'

Таблица 6 – Метрики различных моделей для переменной 'Прочность при растяжении' (полный датасет)

	LinearRegression	ElasticNet	GradientBoosting	RandomForest	KNeighbors	TheilSen
RMSE train	463.955	464.632	453.856	462.280	462.534	464.202
RMSE test	447.200	447.922	447.666	446.289	446.684	446.067
MAE train	370.583	370.830	363.976	368.713	368.818	370.497
MAE test	345.199	345.490	343.169	344.673	345.267	344.981
R2 train	0.009	0.006	0.051	0.016	0.015	0.008
R2 test	0.007	0.004	0.005	0.011	0.009	0.012

3.2.3 'Соотношение матрица - наполнитель'

```
MSE train:  0.763
MSE test:   0.727

RMSE train: 0.873
RMSE test:  0.853

MAE train:  0.704
MAE test:   0.679

R2 train:   0.035
R2 test:    0.026
```

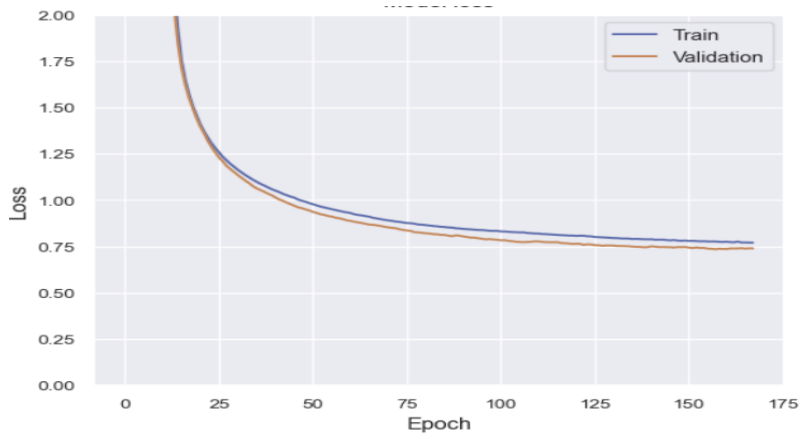


Рисунок 20 – Метрики и график функции потерь полносвязной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для значений переменной 'Соотношение матрица - наполнитель' (полный датасет)

Рисунок 21 – Доверительный интервал для прогноза тестовых значений переменной 'Соотношение матрица - наполнитель' (полный датасет)
Примечание. Зеленым цветом показаны границы доверительного интервала

3.3 Результаты обучения моделей на данных очищенного датасета

В таблице 7 приведены метрики для различных обученных моделей для переменной 'Модуль упругости при растяжении'. Максимальное значение коэффициента детерминации R^2 = 0,49 для тестовой выборки имеет модель GradientBoostingRegressor.

В таблице 8 приведены метрики для различных обученных моделей для переменной 'Прочность при растяжении'. Максимальное значение коэффициента детерминации R^2 = 0,737 для тестовой выборки имеет модель RandomForestRegressor.

Следует отметить, что полученные значения коэффициентов детерминации R^2 для обучающих значений модели LinearRegression идентичны соответствующим значениям аналитического коэффициента детерминации R^2 для МНК – см. таблицу 3.

На рисунке 22 приведены метрики для обученной модели искусственной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для переменной 'Соотношение матрица - наполнитель'. Значение коэффициента детерминации для тестовой выборки составило R^2 = 0,227.

Доверительный интервал для прогноза тестовых значений обучающей выборки переменной 'Соотношение матрица - наполнитель' приведен на рисунке 22. Все значения тестовой выборки расположены в пределах доверительного интервала, что позволяет сделать вывод о приемлемой точности обученной модели.

Полученные метрики моделей, обученных на данных очищенных датасетов, позволяют сделать вывод о том, что на этих данных **удалось** построить неконстантные, то есть пригодные для практического применения, модели для целевых переменных. Данный вывод подтверждает, сделанный в 2.2 вывод о принципиальной возможности построения неконстантных моделей для целевых переменных по объектам очищенных датасетов.

3.3.1 'Модуль упругости при растяжении'

Таблица 7 – Метрики различных моделей для переменной 'Модуль упругости при растяжении' (очищенный датасет: N = 108, k_{nn} = 5)

	LinearRegression	ElasticNet	GradientBoosting	RandomForest	KNeighbors	TheilSen
RMSE train	2.235	2.350	1.606	2.094	2.293	2.291
RMSE test	2.085	2.302	1.936	2.209	2.348	2.309
MAE train	1.743	1.882	1.263	1.580	1.779	1.721
MAE test	1.578	1.785	1.358	1.555	1.884	1.681
R2 train	0.361	0.293	0.670	0.439	0.327	0.328
R2 test	0.408	0.278	0.490	0.335	0.249	0.273

3.3.2 'Прочность при растяжении'

Таблица 8 – Метрики различных моделей для переменной 'Прочность при растяжении' (очищенный датасет: N = 76, k_{nn} = 3)

	LinearRegression	ElasticNet	GradientBoosting	RandomForest	KNeighbors	TheilSen
RMSE train	305.220	328.577	13.572	141.713	215.962	310.517
RMSE test	218.242	279.078	215.996	201.014	233.783	221.547
MAE train	250.507	273.713	9.441	107.150	174.451	257.466
MAE test	184.585	224.106	147.108	141.756	189.464	187.071
R2 train	0.397	0.302	0.999	0.870	0.698	0.376
R2 test	0.689	0.492	0.696	0.737	0.644	0.680

3.3.3 'Соотношение матрица - наполнитель'

```
MSE train:  0.495
MSE test:   0.595
RMSE train: 0.703
RMSE test:  0.771
MAE train:  0.567
MAE test:   0.634
R2 train:   0.102
R2 test:    0.227
```

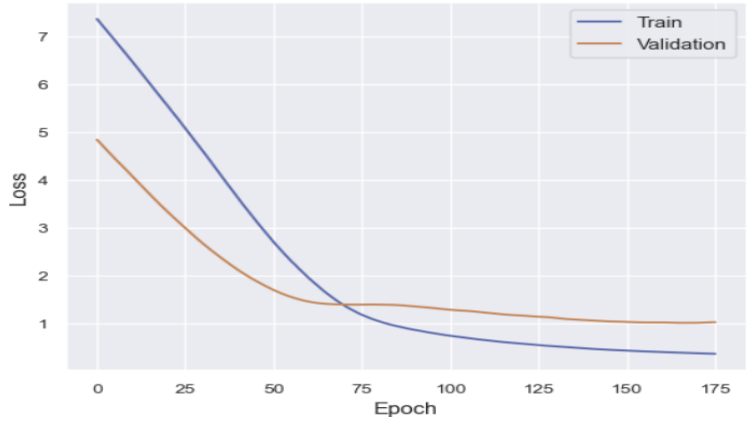


Рисунок 22 – Метрики и график функции потерь полносвязной нейронной сети (2 скрытых слоя с 8 нейронами каждый, 1 выходной слой) для значений переменной 'Соотношение матрица - наполнитель' (очищенный датасет)

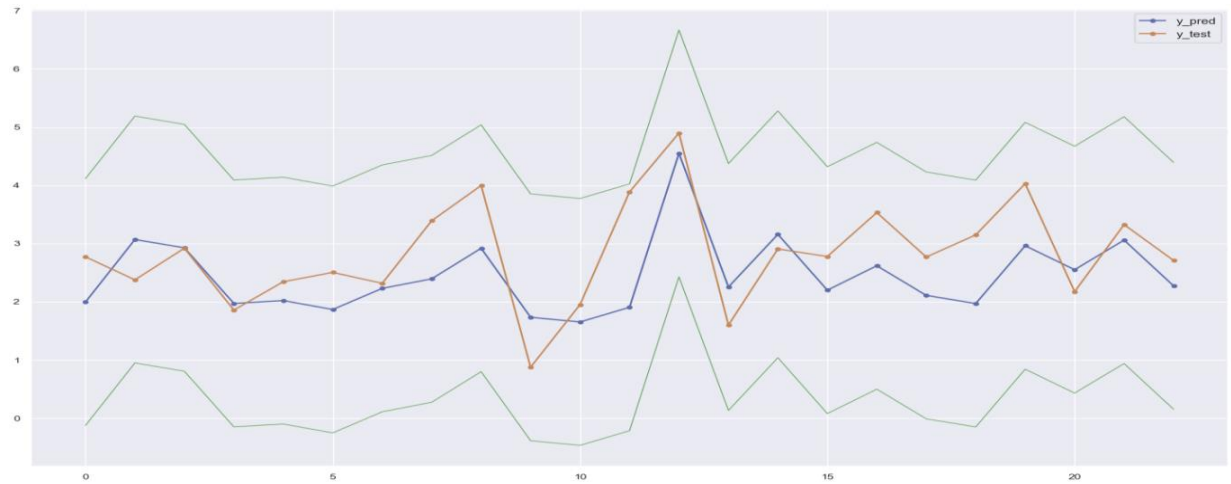


Рисунок 23 – Доверительный интервал для прогноза тестовых значений переменной 'Соотношение матрица - наполнитель' (очищенный датасет)

Примечание. Зеленым цветом показаны границы доверительного интервала

4 СОЗДАНИЕ УДАЛЕННОГО РЕПОЗИТОРИЯ

Результаты ВКР выложены на странице автора на сайте GitHub.com в репозитории BMSTU по адресу https://github.com/GorshkovAndrey/BMSTU

4.1 Дата создания репозитория

Репозиторий создан 08.04.2023.

4.2 Состав репозитория

Репозиторий состоит из трех папок:

1. Для защиты ВКР

В состав папки входят следующие файлы:

first_23.ipynb

y_1_full.ipynb

y_1_filtered.ipynb

y_2_full.ipynb

y_2_filtered.ipynb

y_3_full_NN.ipynb

y_3_filtered_NN.ipynb

y_3_filtered_Regressor.ipynb

Пояснительная записка ВКР. Горшков А.В..docx

Презентация ВКР. Горшков А.В..pptx

2. Первоначальные

В состав папки входят следующие файлы:

y_1.ipynb

y_2.ipynb

y_3.ipynb

3. Фильтрация шума

В состав папки входят следующие файлы:

y_1_Lazy_RdePRk5.ipynb

y_2_Lazy_RdePRk3.ipynb

y_3_Lazy_RdePRk3.ipynb

ЗАКЛЮЧЕНИЕ

1. Модели целевых переменных построены на трех вариантах наборов данных – базовый датасет (первые 23 строки), полный датасет (1023 строки) и очищенный от шума и помех датасет.

2. Для очистки датасета от шума и помех разработан фильтр, ядром (маской) которого являются объекты базового датасета.

3. Для оценки принципиальной возможности построения неконстантной математической модели использовался аналитический коэффициент детерминации для МНК, вычисляемый по коэффициентам корреляции целевых переменных и признаков датасета.

В результате проведенных расчетов аналитического коэффициента детерминации сделан вывод о принципиальной невозможности или малой вероятности построения неконстантных моделей с обобщающей способностью для целевых переменных по объектам полного датасета.

4. По данным базового датасета выявлены линейные функциональные зависимости для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' от признаков 'Поверхностная плотность, г/м2' и 'Потребление смолы, г/м2'. Для целевой переменной 'Соотношение матрица - наполнитель' обучены модели различных регрессоров, для которых наибольший коэффициент детерминации составил R^2 = 0,505.

5. Для полного датасета для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' обучены модели регрессоров, для которых наибольший коэффициент детерминации составил R^2 = 0,012. Для целевой переменной 'Соотношение матрица - наполнитель' обучена модель искусственной нейронной сети, для которой коэффициент детерминации составил R^2 = 0,026.

Полученные метрики моделей, обученных на данных полного датасета, позволяют сделать вывод о том, что на этих данных **не удалось** построить неконстантные модели с обобщающей способностью. Данный вывод

подтверждает приведенный выше вывод о принципиальной невозможности или малой вероятности построения неконстантных моделей для целевых переменных по объектам полного датасета.

6. Для очищенных датасетов для целевых переменных 'Модуль упругости при растяжении' и 'Прочность при растяжении' обучены модели различных регрессоров, для которых наибольший коэффициент детерминации составил $R^2 = 0{,}737$. Для целевой переменной 'Соотношение матрица - наполнитель' обучена модель искусственной нейронной сети, для которой коэффициент детерминации составил $R^2 = 0{,}227$.

Полученные метрики моделей, обученных на данных очищенных датасетов, позволяют сделать вывод о том, что на этих данных **удалось** построить неконстантные, то есть пригодные для практического применения, модели с обобщающей способностью. Данный вывод подтверждает приведенный выше вывод о принципиальной возможности построения неконстантных моделей для целевых переменных по объектам очищенных датасетов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on.

2. Yeo, In-Kwon; Johnson, Richard A. (2000). "A New Family of Power Transformations to Improve Normality or Symmetry". Biometrika. 87 (4): 954–959.

3. Вентцель Е.С. Теория вероятностей: Учеб. для вузов. – 6-е изд. стер. – М.: Высш. шк., 1999

4. Theil. A rank-invariant method of linear and polynomial regression analysis. I, II, III // Nederl. Akad. Wetensch., Proc.. — 1950. —Т. 53. — С. 386–392, 521–525, 1397–1412.